CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Universidade de Brasília: Portal de Periódicos da UnB

RICI Revista Ibero-Americana de Ciência da Informação
ISSN 1983-5213

AFONSO, Alexandre Ribeiro. **B2**: **Um sistema para indexação e agrupamento de artigos científicos em português brasileiro utilizando computação evolucionária**. Brasília, 2013. 157 f., il. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília.

URL: http://repositorio.unb.br/handle/10482/15480

RESUMOS

Resumo: Nesta tese é apresentado um estudo estatístico sobre o agrupamento automático de artigos científicos escritos em português do Brasil, são propostos novos métodos de indexação e agrupamento de textos com o objetivo futuro de desenvolver um software para indexar e agrupar textos por área de conhecimento. Foram testadas três classes conhecidas de termos simples para representar (indexar) os textos de entrada a agrupar: (substantivos), (substantivos e adjetivos), (substantivos, adjetivos e verbos) e também foram desenvolvidas três novas classes de termos compostos para representação (indexação) dos textos: classes de termos mais complexos, onde um termo pode ser composto pela junção de substantivos, adjetivos e preposições. Durante a fase de agrupamento textual dos experimentos foram testados os algoritmos de agrupamento: Expectation-Maximization (EM), X-Means, um Algoritmo Evolucionário de Agrupamento Convencional e, ainda, um novo Algoritmo Evolucionário de Agrupamento Proposto cujo diferencial é trabalhar em duas etapas de processamento: uma etapa para localização do agrupamento subótimo genérico e outra etapa para melhorar tal solução. Adicionalmente, o novo algoritmo permite ao usuário definir a formação de mais grupos ou menos grupos no resultado de agrupamento. Os algoritmos de indexação e agrupamento propostos foram codificados e implementados em um protótipo denominado B2, no entanto, para testar os algoritmos de agrupamento EM e X-Means foi utilizado o pacote de mineração de dados WEKA. Quatro corpora de artigos científicos, diferentes entre si por guardarem artigos de áreas científicas distintas, foram reunidos para testar as combinações de indexação e algoritmo de agrupamento propostas. Melhores resultados de agrupamento (por área de conhecimento dos artigos) foram obtidos utilizando termos compostos na indexação, ao invés do uso de termos simples, quando combinados com o uso do novo Algoritmo Evolucionário de Agrupamento Proposto, porém, para obter grupos bem formados, um número excessivo de grupos é gerado pelo protótipo, consumindo alto tempo de computação para executar tais novos métodos, em um computador pessoal convencional do ano de 2012. Pode-se concluir que o problema de agrupar automaticamente artigos científicos em suas áreas originais é uma tarefa complexa. Logo, acredita-se que os métodos de indexação e agrupamento desenvolvidos possam ser aprimorados para utilização futura em situações específicas, onde a fragmentação e geração adicional de grupos além do esperado não seja um problema maior.
Palavras-chave: Algoritmos; Artigo científico; Indexação automática; Linguística - processamento de dados; Mineração de texto.

Abstract: This thesis presents an empirical study about automated text clustering for scientific articles written in Brazilian Portuguese. We tested three already known classes of simple terms for representing (or indexing) the input texts: (nouns), (nouns and adjectives) and (nouns, adjectives and verbs); we also developed three new classes of composed terms for text representation (or indexing): the new classes consist of more complex terms, where a complex term could be composed by the joint of nouns, adjectives and prepositions. Our final goal is to develop new software for text indexing and clustering. During the clustering stage of the experiments we tested the Expectation-Maximization (EM) Clustering Algorithm, the X-Means Clustering Algorithm, the Conventional Clustering Evolutionary Algorithm and, finally, we also proposed a new Two Phase Clustering Evolutionary Algorithm which works in two phases, the first phase finds the sub-optimal text clustering and the second one improves the result found by the first phase. The Two Phase Clustering Evolutionary Algorithm also permits the user to define whether the system should create a high number or a low number of clusters. The new indexing and clustering algorithmic strategies presented were implemented in a prototype named B2, but for testing the EM and X-Means algorithms we used the known WEKA data mining package. Four different scientific corpora having different sets of scientific topics were assembled and applied for testing the combinations of indexing and clustering methods. Although considerable better results were achieved when indexing with the classes of composed terms combined with the new Two Phase Clustering Evolutionary Algorithm, a considerable higher number of clusters was generated and a considerable additional time was consumed when running the new system over a 2012 conventional personal computer. We conclude that the problem of clustering scientific articles in their original topics is a complex task. Good results of clustering correctness were achieved by the new methods but producing many fragmented additional clusters as output, so, in the future, the methods can be improved and applied in specific situations where the fragmentation and additional production of clusters are not a major problem.
Keywords: Algorithm; Automatic indexing; Data mining; Linguistics – data processing; Scientific article; Text mining.